Self-Organization in Decentralized Networks: A Trial and Error Learning Approach

Luca Rose, *Student Member, IEEE,* Samir M. Perlaza, *Member, IEEE,* Christophe J. Le Martret, *Senior Member, IEEE,* and Mérouane Debbah, *Senior Member, IEEE*

***Abstract*—In this paper, the problem of channel selection and power control is jointly analyzed in the context of multiple-channel clustered ad-hoc networks, i.e., decentralized networks in which radio devices are arranged into groups (clusters) and each cluster is managed by a central controller (CC). This problem is modeled by game in normal form in which the corresponding utility functions are designed for making some of the Nash equilibria (NE) to coincide with the solutions to a global network optimization problem. In order to ensure that the network operates in the equilibria that are globally optimal, a learning algorithm based on the paradigm of trial and error learning is proposed. These results are presented in the most general form and therefore, they can also be seen as a framework for designing both games and learning algorithms with which decentralized networks can operate at global optimal points using only their available local knowledge. The pertinence of the game design and the learning algorithm are highlighted using specific scenarios in decentralized clustered ad hoc networks. Numerical results confirm the relevance of using appropriate utility functions and trial and error learning for enhancing the performance of decentralized networks.**

***Index Terms*—Ad-hoc networks, Resource allocation, Interference management, QoS provisioning, Game theory**

I. Introduction

A decentralized self-configuring network (DSCN) is an infrastructure-less network in which transmitters communicate with their respective receivers without the control of a central authority, for instance, a base station. The relevance of these networks lies on the fact that a formal network planning is not required, their deployment is easy, quick and, more importantly, capabilities such as self-healing and self-configuration are often present. Therefore, DSCNs span a large number of applications including military, law enforcement, disaster relief, space, and indoor/outdoor commercial applications [1], [2].

A growing body of research suggests that self-organization is one of the fundamental capabilities decentralized networks must exhibit [3]–[8]. The term self-configuration refers to the capability of radio devices to autonomously tune their transmit-receive configuration for efficiently exploiting the available resources and guaranteeing network reliability. In the most general case, a transmit-receive configuration can be described in terms of the number of information bits per block, the block length, the codebook, the encoding-decoding functions, the channel selection policy, the power allocation policy, etc., as suggested in [9]–[11].

This paper focuses exclusively on the self-configuration dimension of these networks and more specifically, on the case of multiple-channel clustered ad hoc networks, i.e., DSCNs in which radio devices are arranged into groups (clusters) and each cluster is managed by a central controller (CC). The main task of the CC is to choose the logical channel in which its cluster must operate and to determine the power levels to be used by all radio devices inside the cluster. Hence, this network model is both decentralized, in the sense that there exist several CCs autonomously taking decisions, and also centralized, in the sense that radio devices inside a cluster implement the decision adopted by their corresponding CC. The decision-taking problem faced by all CCs is modeled by a game in normal form and the corresponding utility functions are designed for making some of the Nash equilibria (NE) to coincide with the solutions to a global network optimization problem. That is, the utility functions are designed to make some of the equilibria to be global optimal operating points. In order to ensure that the network operates in those equilibria that are globally optimal, a learning algorithm based on the paradigm of trial and error learning [12] is proposed. The main characteristics of the proposed algorithm are studied and interesting conclusions in terms of time to reach a globally optimal NE are presented. Interestingly, the results presented here are introduced in the most general form in order to provide not only a solution to the problem described above but also to provide a general framework for designing both games and learning algorithms with which decentralized networks can operate at global optimal points using only their available local knowledge about the network.

The use of game theory [13] and multi-agent learning [12] is already well accepted in the analysis of this kind of problems. The following sub-section highlights the most relevant contributions in this sense.

A. State of the Art

Among the most relevant contributions regarding the main results of this paper, it is worth to highlight those in [14]–

Manuscript received March 5, 2012; revised September 5, 2013; accepted October 13 2013.

L. Rose is with the Alcatel-Lucent chair in Flexible radio, Supélec, Plateau du Moulon, 91192, Gif-sur-Yvette, France and with Thales Communications & Security, avenue des Louvresses, 92622, Gennevilliers, France (e-mail: (luca.rose@supelec.fr).

S.M. Perlaza is with the School of Engineering and Applied Science at Princeton University, New Jersey, USA. (e-mail: perlaza@princeton.edu).

C. J. Le Martret is with Thales Communications & Security, avenue des Louvresses, 92622, Gennevilliers, France (e-mail: christophe.le_martret@thalesgroup.com).

M. Debbah is with the Alcatel-Lucent chair in Flexible radio, Supélec, Plateau du Moulon, 91192, Gif-sur-Yvette, France (e-mail: merouane.debbah@supelec.fr).

[23]. In [14], the authors consider a clustered multi-channel ad hoc network in which clusters are able to sense all available channels in order to choose an interference-free channel. When an interference-free channel is not available, the choice on the channel is randomly made. In low population density networks, this behavioral rule is shown to exhibit an acceptable performance with very little implementation complexity. Nonetheless, in high population-density networks, this approach is also shown to be highly suboptimal. Other approaches are based on the use of the iterative water-filling (IWF) power allocation policy [15]–[17]. For instance, in [15], the authors prove that the IWF algorithm converges to an operating point that guarantees a given transmission rate while minimizing the transmit power. Such a convergence is subject to the assumption that the system operates in the weak interference regime. In [24]–[26], it is shown that in decentralized networks the operating point achieved by using IWF is often inefficient. To overcome this inefficiency, solutions based on the cooperation of all transmitters are proposed in [27]–[29]. Among other approaches, including those based on reinforcement learning, maximum-entropy reinforcement learning, smoothed best-response or fictitious play, it is important to highlight the contributions in [3], [7], [8], [18]–[23]. The main drawbacks of these contributions can be summarized in five points: (i) The converging point is a probability distribution over the set of all available channels and power allocations policies [21], [22], [30], [31]. Therefore, the optimization is often on the expectation of the performance metric and the optimality is often claimed in the asymptotic regime. (ii) Often, optimal performance is claimed only in the case in which the number of available channels is higher than the number of devices [20]; (iii) The results are not general as often algorithms are designed for a particular metric, e.g., the transmission rate [4], [7], [8]; (iv) The channel selection problem is treated separately from the power allocation problem [19], [20]; or (v) only the power control problem is considered [23].

B. Structure of the Paper

The rest of the paper unfolds as follows. Sec. II introduces the system model and describes a game in normal form that is used as reference to present the main results. Sec. III briefly describes the trial and error algorithm as introduced in [32]. Sec. IV establishes a connection between the stochastically stable points of the TE algorithm, the NE of the game cited above and the solutions to a global network optimization problem. Sec. V presents some simulation results in oder to highlight some heuristic improvements that can be implemented to the classical TE algorithm. Sec. VI concludes this work.

II. Problem Formulation

This section describes the network model and a centralized network optimization problem whose solutions are assumed to be the desirable network operating points. Later, this section introduces a game in normal form whose utility function is designed for making its Nash equilibria to coincide with the solutions of the network optimization problem.

A. System Model

Consider a DSCN in which all its nodes coexist within the same spectrum subject to mutual interference. Nodes are arranged into groups, referred to as cells. Each cell is controlled by a CC that harmonizes the intra-cell communications by strategically choosing a channel (frequency band) and a power level to be used by all the nodes in the corresponding cell. This model represents either networks in which all nodes are interested in communicating with the same receiver (e.g., the CC acts also as a receiver) or networks in which the CC manages several point-to-point communications inside the cell. Let $\mathcal{K} = \{1, 2, \ldots, K\}$ be a set of K cells. Let also $\mathcal{L}_k = \{\ell_{1,k}, \ell_{2,k}, \ldots, \ell_{k,L_k}\}$ denote the set of L_k links within cell k, with $k \in \mathcal{K}$. The set of all the links in the network is denoted by $\mathcal{L} = \cup_{k\in\mathcal{K}}\mathcal{L}_k$, with $L = |\mathcal{L}|$ the total number of links in the network.

Let $\mathcal{C} = \{1, 2, \ldots, C\}$ be the set of C channels into which the total spectrum is divided. All channel gains are assumed to be time invariant for the whole duration of one transmission. Cell k uses only one channel denoted by $c_k \in \mathcal{C}$ and a transmit power level p_k that is chosen from a finite set $\mathcal{P} = \{0, \ldots, P_{\max}\}$ of $Q = |\mathcal{P}|$ power levels. The maximum transmittable power level is denoted by $P_{\max}$ and it is assumed to be the same for all cells. A pair of a channel and a power level is referred to as an *action*, i.e, $a_k = (c_k, p_k) \in \mathcal{A}$, where $\mathcal{A} = \mathcal{C} \times \mathcal{P}$ is the set of actions. The super vector describing the whole network configuration is denoted by $\boldsymbol{a} = (a_1, a_2, \ldots, a_K) \in \mathcal{A} \times \ldots \times \mathcal{A} = \mathcal{A}^K$, and it is often referred to as an action profile.

The goal is to design a fully decentralized algorithm that selects a network configuration vector $\boldsymbol{a}^* \in \mathcal{A}^K$ that is a solution of the following optimization problem

$$\begin{cases} \max\limits_{\boldsymbol{a}\in\mathcal{A}^K} \sum\limits_{k=1}^{K} \varphi_k(\boldsymbol{a}) \\ \text{s.t. } \xi_\ell(\boldsymbol{a}) > \Gamma \qquad \forall \ell \in \mathcal{L}^*. \end{cases} \tag{1}$$

The function $\varphi_k : \mathcal{A}^K \to [0, 1]$ determines the performance $\varphi_k(\boldsymbol{a})$ achieved by the cell k when the actions chosen by all cells correspond to the action profile $\boldsymbol{a}$. The function $\xi_\ell(\cdot) : \mathcal{A}^K \to [0, 1]$ represents the QoS constraints to which link ℓ is subject. The set $\mathcal{L}^* \subseteq \mathcal{L}$ is defined as the largest set of links for which the constraints in (1) can be simultaneously satisfied. Note that $\mathcal{L}^*$ depends on all the individual constraints that are autonomously determined by each link. Thus, not all the constraints might be simultaneously satisfiable. Fixing the set $\mathcal{L}^*$ is a mathematical maneuver in order to guarantee that the optimization domain in (1) is not empty. Later, it is shown that there is no loss of generality by assuming the set $\mathcal{L}^*$ to be known in advance. The formulation in (1) might describe a large set of network optimization problems that do not necessarily need to be convex. For instance, by properly selecting the functions φ_k and ξ_ℓ, it is possible to analyze problems such as: (a) the throughput maximization problem subject to particular delay constraints; (b) the transmit power minimization subject to a particular network reliability constraint; and other problems.

Here, the final goal is to design a decentralized behavioral rule that allows the network to achieve an operating point $\boldsymbol{a}^*$ that is a solution of (1) based only on local intra-cell available information.

B. Game Theoretical Modeling

This sub-section introduces a game in normal form

$$\mathcal{G} = \left(\mathcal{K}, \{\mathcal{A}_k\}_{k\in\mathcal{K}}, \{u_k\}_{k\in\mathcal{K}}\right), \tag{2}$$

that models the problem described in Sec. II. The set $\mathcal{K}$ represents the players, i.e., the K central controllers in the network; the set $\mathcal{A}$ represents the individual actions of all players. Note that all players have the same set of actions. An action of player k, denoted by $a_k = (c_k, p_k) \in \mathcal{A} = \mathcal{C}\times\mathcal{P}$, is a pair made of a channel index (frequency band) and the transmit power level to be used by all links inside the corresponding cell. The utility function of player k is $u_k : \mathcal{A}^K \to [0,1]$ and it is defined as

$$u_k(\boldsymbol{a}) = \frac{1}{1+\beta L_{\max}}\left(\varphi_k(\boldsymbol{a}) + \beta\sum_{\ell\in\mathcal{L}_k}\mathbb{1}_{\{\xi_\ell(\boldsymbol{a})>\Gamma\}}\right) \tag{3}$$

where β is a design parameter that tunes the tradeoff between the number of links that can be satisfied $\sum_{\ell\in\mathcal{L}_k}\mathbb{1}_{\{\xi_\ell(\boldsymbol{a})>\Gamma\}}$, and the maximization of the function φ_k. The utility function in (3) is a generalization of the utility function proposed in [33]. More importantly, it is chosen such that it exhibits several useful properties. First, it is monotonically increasing with both the number of links that are able to satisfy their individual constraints inside the corresponding cell k; and with the value of the function φ_k that determines the global performance of cell k. Second, as shown in Sec. IV, for a particular choice of the parameter β, there exists an algorithm whose stochastically stable points are both equilibria of the game $\mathcal{G}$ and solutions of the optimization problem in (1).

The notion of equilibrium used in the following of this analysis is that proposed by Nash in [34].

Definition 1 *(Nash equilibrium in pure strategies). An action profile $\boldsymbol{a}^* \in \mathcal{A}^K$ is an NE of the game $\mathcal{G}$ if $\forall k \in \mathcal{K}$ and $\forall a'_k \in \mathcal{A}$*

$$u_k(a^*_k, \boldsymbol{a}^*_{-k}) \geq u_k(a'_k, \boldsymbol{a}^*_{-k}). \tag{4}$$

The motivation for adopting the NE as the solution concept of the game $\mathcal{G}$ relies on the fact that at an NE, the action adopted by every player is optimal with respect to the choices made by all the other players. That is, if a player k decides to deviate from its individual NE action, its utility can only be degraded if the system is at such NE. Therefore, from a decentralized point of view, this property is particularly desirable.

In the case the game $\mathcal{G}$ possesses several equilibria, they can be compared by calculating their corresponding sum-utility, i.e., the sum of the utilities obtained by all players at the corresponding equilibrium. Therefore, the rest of this analysis focuses on those NE that are maximizers of the *social welfare* function $W : \mathcal{A}^K \to \mathbb{R}$ defined by $W(\boldsymbol{a}) = \sum_{k=1}^K u_k(\boldsymbol{a})$.

III. Trial and Error Learning Algorithms

The purpose of this section is threefold. First, it provides a brief description of the trial and error (TE) algorithm as introduced in [12] and [32]; second, it describes the adaptation of this algorithm to the game described in Sec. II-B; finally, it introduces some heuristic enhancements to optimize the convergence properties.

A. Trial and Error Learning Algorithm

The TE learning algorithm can be described by a state machine locally implemented by each player. The main feature of this state machine is that the set of stochastically stable states are the NE that maximize the social welfare.

At each iteration t, the state of player k is defined by the triplet:

$$Z_k(t) = \{m_k(t), \bar{a}_k(t), \bar{u}_k(t)\}, \tag{5}$$

where $m_k(t) \in \{C, C^+, C^-, D\}$ represents the *mood*: *content* (C), *hopeful* (C^+), *watchful* (C^-), *discontent* (D), $\bar{a}_k(t) \in \mathcal{A}$ and $\bar{u}_k(t) \in [0,1]$ represent the *benchmark* action and *benchmark* utility, respectively. The state machine transitions and behavior are detailed hereunder. Note that the notation $a \Leftarrow b$ indicates that variable a takes the value of variable b.

Content: Let $\epsilon \in [0,1]$ be an experimentation parameter and assume that the state of player k at time $t-1$ is $Z_k(t-1) = \{C, \bar{a}_k(t-1), \bar{u}_k(t-1)\}$. Then, at iteration t, it selects its action according to the following rule: with probability $(1-\epsilon)$, it plays the benchmarked action $a_k(t) = \bar{a}_k(t-1)$ or with probability ϵ, it plays another action randomly selected $a_k(t) \neq \bar{a}_k(t-1)$. Once player k has played action $a_k(t)$, it observes the value of its utility function $u_k(t)$.

The player updates its state as follows: If $a_k(t) \neq \bar{a}_k(t-1)$ and $u_k(t) \leq \bar{u}_k(t-1)$, then $Z_k(t) \Leftarrow Z_k(t-1)$; If $a_k(t) \neq \bar{a}_k(t-1)$ and $u_k(t) > \bar{u}_k(t-1)$, then, with probability $\epsilon^{G(u_k(t)-\bar{u}_k(t-1))}$, it sets $m_k(t) \Leftarrow m_k(t-1)$, $\bar{a}_k(t) \Leftarrow a_k(t)$ and $\bar{u}_k(t) \Leftarrow u_k(t)$, while with probability $\left(1-\epsilon^{G(u'_k(t)-\bar{u}_k(t-1))}\right)$, it sets $Z_k(t) \Leftarrow Z_k(t-1)$; If $a_k(t) = \bar{a}_k(t-1)$ and $u_k(t) \geq \bar{u}_k(t-1)$ then, $m_k(t) \Leftarrow C^+$, $\bar{a}_k(t) \Leftarrow \bar{a}_k(t-1)$, $\bar{u}_k(t) \Leftarrow \bar{u}_k(t-1)$; If $a_k(t) = \bar{a}_k(t-1)$ and $u_k(t) < \bar{u}_k(t-1)$ then $m_k(t) \Leftarrow C^-$, $\bar{a}_k(t) \Leftarrow \bar{a}_k(t-1)$, $\bar{u}_k(t) \Leftarrow \bar{u}_k(t-1)$.

Note that if player k does not experiment (it plays its benchmarked action) and its utility increases, then it becomes *hopeful*, while if it decreases, it becomes *watchful*. Here, the function $G : \mathbb{R} \to \mathbb{R}$ must be such that:

$$0 \leq G(x) < \frac{1}{2}. \tag{6}$$

Numerical simulations suggest that a linear formulation such as: $G(\Delta u) = -0.2\Delta u + 0.2$, with $\Delta u = u_k(t) - \bar{u}_k(t-1)$, performs well under several scenarios.

Hopeful: Assume that the state of player k at time $t-1$ is $Z_k(t-1) = \{C^+, \bar{a}_k(t-1), \bar{u}_k(t-1)\}$. Then, at iteration t, it plays the benchmark action $a_k(t) = \bar{a}_k(t-1)$ and it observes the value of its utility function $u_k(t)$. If $u_k(t) \geq \bar{u}_k(t-1)$ then, $m_k(t) \Leftarrow C$, $\bar{a}_k(t) \Leftarrow \bar{a}_k(t-1)$ and $\bar{u}_k(t) \Leftarrow \bar{u}_k(t-1)$; otherwise, $m_k(t) \Leftarrow C^-$, $\bar{a}_k(t) \Leftarrow \bar{a}_k(t-1)$ and $\bar{u}_k(t) \Leftarrow \bar{u}_k(t-1)$.

Watchful: Assume that the state of player k at time $t-1$ is $Z_k(t-1) = \{C^-, \bar{a}_k(t-1), \bar{u}_k(t-1)\}$. Then, at iteration t, it plays the benchmark action $a_k(t) = \bar{a}_k(t-1)$ and it observes the value of its utility function $u_k(t)$. If $u_k(t) > \bar{u}_k(t-1)$, then $m_k(t) \Leftarrow C^+$, $\bar{u}_k(t) \Leftarrow \bar{u}_k(t-1)$ and $\bar{a}_k(t) \Leftarrow \bar{a}_k(t-1)$; otherwise, $m_k(t) \Leftarrow D$, $\bar{u}_k(t) \Leftarrow \bar{u}_k(t-1)$ and $\bar{a}_k(t) \Leftarrow \bar{a}_k(t-1)$.

Discontent: Assume that the state of player k at time $t-1$ is $Z_k(t-1) = \{D, \bar{a}_k(t-1), \bar{u}_k(t-1)\}$. Then, at iteration t, it randomly selects an action $a_k(t)$ and observes the value of its utility function $u_k(t)$. The state is updated as follows: with probability $p = \epsilon^{F(u_k(t))}$ it sets $m_k(t) \Leftarrow C$, $\bar{u}_k(t) \Leftarrow u_k(t)$ and $\bar{a}_k(t) \Leftarrow \bar{a}_k(t-1)$; with probability $(1-p)$ it sets $m_k(t) \Leftarrow D$, $\bar{u}_k(t) \Leftarrow u_k(t)$ and $\bar{a}_k(t) \Leftarrow a_k(t)$. The function $F : \mathbb{R} \to \mathbb{R}$ must be such that

$$0 \leq F(u) < \frac{1}{2K}. \tag{7}$$

Numerical simulations suggest that a linear formulation such as: $F(u) = -\frac{0.2}{K}u + \frac{0.2}{K}$ performs well under several scenarios.

In [12] and [32], the authors proved that the stochastically stable action profiles of the trial and error algorithm (i.e., action profiles that are played most of the time) are those NE that maximize the social welfare. Theorem 1 restates their main results.

Theorem 1 *Let $\mathcal{G}$ have at least one pure NE and let each player use TE. Then, for each ϵ small enough there exists a δ such that a pure Nash equilibrium that maximizes the sum utility among all equilibrium states is played $(1-\delta)$ fraction of the time.*

Theorem 1 states that if all players implement the TE algorithm and there exists at least one NE, then the NE with the highest social welfare is played during a *large* fraction of the time. In general, the quantity $1-\delta$ depends on ϵ and on the particular game $\mathcal{G}$. When players implement the TE algorithm, the notion of convergence largely differs from the classical idea of convergence, that is, a dynamic distance minimization with respect to a certain action profile (e.g., an NE, a correlated equilibria, etc), indeed, once such a limiting action profile is reached, the system remains static. The convergence of the TE algorithm must be understood in terms of the time players remain at a given action profile. Indeed, the system can be at an NE, but it might arbitrarily leave it to experiment other action profiles. Therefore, in this setting, convergence refers to the fact that the system remains on certain action profiles a large fraction of the time.

B. Enhanced Distribution and Settings

This section presents some enhancements to the TE algorithm in order to improve its performance. In its standard formulation, the TE learning algorithm [12] is characterized by a time invariant ϵ and a uniform distribution over the whole action set. Motivated by the fact that experimentations on the set of channels brings higher instability than experimentations on the set of power levels, the experimentation is divided into two different steps. In detail, at each instant t, each player in a *content* mood and denoted by k experiments with probability $\epsilon_c^k(t)$ a different channel and with probability $\epsilon_p^k(t)$ a different power level. The evolution of $\epsilon_c^k(t)$ is given by:

$$\begin{cases} \epsilon_c^k(t) &= \max\left(\frac{\epsilon_c^k(t-1)}{2}, \epsilon_c^{min}\right) & \text{if } \sum_{n\in\mathcal{L}_k} \mathbb{1}_{\{\phi_n(\boldsymbol{a})>\Gamma\}} = |\mathcal{L}_k| \\ \epsilon_c^k(t) &= \epsilon_c^k(0) & \text{otherwise.} \end{cases} \tag{8}$$

In (8), $\epsilon_c^{min} > 0$ represents the minimum experimentation probability over the available channels and $\epsilon_c^k(0) > \epsilon_c^{min}$ represents the initial value. These parameters depend on the particular configuration of the system. Through numerical simulations, it has been found that some well-performing values are: $\epsilon_c^{min} = \frac{0.01}{K}$ and $\epsilon_c^k(0) = 0.01\frac{C}{K}$. Since no prior information is available on the channels' gain, the experimentation on the channels follows a uniform distribution.

Each player k experiments a different power level with a constant probability ϵ_p^k. Such a probability is a uniform distribution over all the levels greater than p_k if $\sum_{n\in\mathcal{L}_k} \mathbb{1}_{\{\phi_n(\boldsymbol{a})>\Gamma\}} < |\mathcal{L}_k|$, whereas it is uniformly distributed over all the levels smaller than p_k, otherwise. Through extensive simulations, it has been found that a well-performing value is $\epsilon_p^k = 0.01\frac{C}{K}$.

When a player k is *discontent*, it experiments according to the following distribution:

$$\begin{cases} p_k(t) &= P_{\max} & \text{with probability } \min\left(\frac{C}{K}, 1\right) \\ p_k(t) &= 0 & \text{with probability } \max\left(1-\frac{C}{K}, 0\right) \end{cases}. \tag{9}$$

The rationale behind this is that any *discontent* player needs to test the network looking for a free channel. Clearly, the probability of finding a free channel increases with $\frac{C}{K}$. On the other hand, in the case in which no channel is free for transmission, zero power should be used to avoid wasting energy and creating interference.

IV. Convergence study

This section presents the main theoretical results of the paper. A strong connection between the solutions of the optimization problem in (1) and the NE of the game $\mathcal{G}$ is established via the utility function (3).

A. Equilibrium points

Theorem 2 *Let all the players of the game $\mathcal{G}$ implement the TE algorithm with the utility function in* (3)*, and let $\beta \in \mathbb{R}$ satisfy $\beta > K$. Let also $\mathcal{A}_{\text{NE}}$ be the set of NE of the game $\mathcal{G}$ and assume it is non empty. Denote by λ_n the number of links satisfied at the n-th NE, with $n \in \{1, \ldots, |\mathcal{A}_{\text{NE}}|\}$ and let $\Lambda = \max_{n\in\{1,\ldots,|\mathcal{A}_{\text{NE}}|\}} \lambda_n$. Then, the TE algorithm is stochastically stable in an NE in which there are at least Λ links that satisfy their individual constraints.*

Theorem 2 states that, if each player sets $\beta > K$, then the stochastically stable points of the TE learning algorithm are those NE with the largest possible number of links satisfying their constraints. Here, β represents the trade-off between the interest in satisfying the constraints for the largest set of links and the maximization of the sum of the objective functions.

The next theorem links this result with the global optimization problem in (1).

Theorem 3 *Let all the players of the game $\mathcal{G}$ implement the TE learning algorithm with the utility function in* (3)*, and let $\beta \in \mathbb{R}$ satisfy $\beta > K$. Let $\mathcal{A}^\dagger \subseteq \mathcal{A}^K$ be the set of solutions of the optimization problem in* (1)*, and let $\mathcal{L}^*$ be the largest set such that $\exists\ \boldsymbol{a} \in \mathcal{A}^\dagger$ and $\forall \ell \in \mathcal{L}^*,\ \xi_\ell(\boldsymbol{a}) > \Gamma$ and $|\mathcal{L}^*| = L^*$. Let also $\mathcal{A}_{\rm NE}$ be the set of NE of the game $\mathcal{G}$, and assume $\mathcal{A}_{\rm NE} \cap \mathcal{A}^\dagger$ is non-empty. Then, the TE algorithm is stochastically stable in an action profile $\boldsymbol{a}^*$ such that $\boldsymbol{a}^* \in \mathcal{A}_{\rm NE} \cap \mathcal{A}^\dagger$.*

Note that the set of solutions of (1) is non-empty as long as there exists a set $\mathcal{L}^*$ such that the optimization domain is not an empty set. This theorem states that the stochastically stable points of the TE algorithm are those NE that maximize the sum of the network objective functions among the action profiles that satisfy the constraints for the largest possible set of links. For instance, if the network objective functions $\varphi_k(\cdot)$ are decreasing with respect to the power level p_k, then the stochastically stable points are those NE which satisfy the constraints for the largest number of links and minimize the power consumption.

B. Convergence time

This section studies the convergence properties of the TE algorithm in a particular scenario.

More specifically, the TE learning algorithm defines a large discrete time Markov chain (DTMC) over the set of the states. Studying the behavior of the algorithm on such a chain is a difficult problem due to the number of states, transitions and parameters. For this reason, a simplified version of the system model introduced in Sec. II-B is considered. This allows the estimation of the average number of time instants that are required to reach an NE for the first time and the expected fraction of time the system is at an NE action profile.

For the ease of the presentation, consider $L_k = 1$, i.e., each cell possesses only one link. The functions φ and ξ are defined as:

$$\begin{cases} \varphi_k(\boldsymbol{a}) &= 1 - \frac{p_k}{P_{\max}} \\ \xi_k(\boldsymbol{a}) &= \mathrm{SINR}_k(\boldsymbol{a}). \end{cases} \tag{10}$$

In this particular formulation, the aim is to minimize the transmit power while keeping the SINR above a threshold Γ for the largest number of links. In (10), since there is only one link per cell, the link index is the same as the cell index. Therefore the SINR of link k is evaluated as:

$$\mathrm{SINR}_k(\boldsymbol{a}) = \frac{p_k g_{k,k}^{(c_k)}}{\sigma^2 + \sum\limits_{\ell \in \mathcal{K} \setminus k} p_\ell g_{k,\ell}^{(c_\ell)} \mathbb{1}_{\{c_\ell = c_k\}}}, \tag{11}$$

where $g_{k,\ell}^{(c_k)}$ indicates the channel power gain between the transmitter of link k and the receiver of link ℓ over channel c_k; and σ^2 represents the noise power. This problem has been also studied in [33]. Note that it is possible for the receivers to evaluate the SINR through pilots or training sequences. In the following, it is assumed that $C \geq K$, and that the channel gains follows the weak interference model as in [10]:

$$\begin{cases} g_{k,k}^{(c)} = 1 & \forall k \in \mathcal{K} \text{ and } \forall c \in \mathcal{C} \\ g_{j,k}^{(c)} = \frac{1}{2} & \forall k \in \mathcal{K} \text{ and } \forall j \in \mathcal{K} \setminus \{k\} \text{ and } \forall c \in \mathcal{C} \end{cases}. \tag{12}$$

In the light of the description made in Sec. III, if the number of players K is large enough the following can be stated: (i) the fraction of time player k is either at *watchful* or *hopeful* state is negligible compared to the fraction of time it spends in *discontent* or *content* state; (ii) at any time, the probability of having more than one player *discontent* is significantly lower than the probability of having only one or no *discontent* player. In fact, in (7) the probability of accepting the outcome of the experimentation for a player which is discontent is close to one, moreover players do not adopt a *watchful* or *hopeful* state for more than one iteration. Sec. V shows that these results are good approximations under less restrictive conditions as well.

Under these conditions, the resulting DTMC for studying the TE learning algorithm is represented in Fig. 2.

In this figure, the final state represents an NE, the states labeled with C_{K-k} are those in which $K-k$ players use an individually optimal action and D a state in which one player is *discontent*. The transition probabilities are listed hereafter (the reasoning behind these transition probabilities is given in the appendix):

$$P(N, D) = \frac{K(K-1)^2\epsilon^2}{C^2}\left(\frac{Q-1}{Q}\right)^2 \tag{13}$$

$$P(D, N) = \frac{(C-K+1)}{CQ} \tag{14}$$

$$P(D, C_{K-k}) = \frac{(C-K+k)}{C^k}\frac{(K-1)!}{(K-k)!} \tag{15}$$

$$P\left(C_{K-k}, C_{K-k-1}\right) = (K-k)\frac{C-k}{CQ}\epsilon. \tag{16}$$

Here, $P(N, D)$ is the transition probability between an NE and a state in which one player is *discontent*; $P(D, N)$ is transition

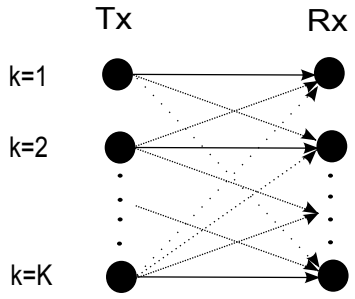


Fig. 1. Simplified system model: symmetric parallel interference channel.

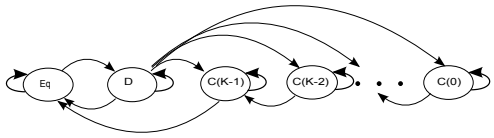


Fig. 2. Markov chain describing the TE learning algorithm in the network. This model is used to study the convergence to the NE. The state Eq represents an NE action profile. C_{K-k} represents a state in which $K-k$ players are using an individually optimal action, D represents a state in which at least one player is *discontent*.

probability between a state in which one player is *discontent* and an NE; $P(D, C_{K-k})$ is the transition probability between a state in which one player is *discontent* and a state in which $K-k$ players are using an individually optimal action; and $P\left(C_{K-k}, C_{K-k-1}\right)$ is the transition probability between a state in which $K-k$ players are using an individually optimal action and a state in which $K-k-1$ are doing the same. The analysis of this DTMC leads to state the following theorems.

Theorem 4 *Let K, C, Q, and ϵ be the number of players, the number of channels, the number of power levels and the experimentation parameter respectively. Assume $C \geq K$. Let $L_k = 1$ and let the channel power gains be given by* (12)*. Then, if all players implement the TE learning algorithm, the expected number of iterations needed to reach the NE for the first time, $\bar{T}_{NE}$, is bounded as follows:*

$$\bar{T}_{NE} \leq \frac{CQ}{\epsilon\left(C-K\right)}\left(1+\log\left(\frac{K\left(C-K+1\right)}{C+1}\right)\right) \quad (17)$$

$$\bar{T}_{NE} \geq \frac{CQ}{\epsilon\left(C-K\right)}\left(\gamma+\log\left(\frac{K\left(C-K\right)}{C}\right)\right); \quad (18)$$

where, $\gamma \simeq 0.577$ is the Euler-Mascheroni constant.

Note that the time needed to visit for the first time an NE is directly proportional to the dimension of the action set (i.e., $|\mathcal{A}| = CQ$) and inversely proportional to the experimentation probability ϵ.

Theorem 5 *Let K, C, Q, and ϵ be the number of players, the number of channels, the number of power levels and the experimentation parameter, respectively. Assume $C \geq K$, $L_k = 1$, and let also the channel power gains follow* (12)*. Then, if all players follow the TE learning algorithm the expected fraction of time the system is at an NE is:*

$$(1-\delta) \approx \frac{1}{1+P(N,D)T_{BNE}}, \quad (19)$$

where

$$T_{BNE} \approx \sum_{k=1}^{K} P(D, C_{K-k}) T_{CNE}(k) + \frac{P(D,N)}{\left(1-P(D,D)\right)^2},$$

$$T_{CNE}(k) \approx \frac{CQ}{\epsilon\left(C-K\right)}\left(\gamma+\log\left(\frac{K\left(C-k+1\right)}{C+1}\right)\right),$$

$$P(D,D) = 1 - P(D,N) - \sum_{k=1}^{K} P(D, C_{K-k}).$$

Note that the frequency of using an NE, i.e., $(1-\delta)$ depends on $\frac{1}{\epsilon^2}$ as in (13). This means that the larger the ϵ the shorter the time the system is at an NE. The approximation is given by the fact that T_{BNE} is replaced by its upper bound.

These theorems show that the stability of the TE algorithm and the time it needs to visit an NE for the first time are influenced by the experimentation probability. Lower values increase stability while higher values increase the speed of convergence.

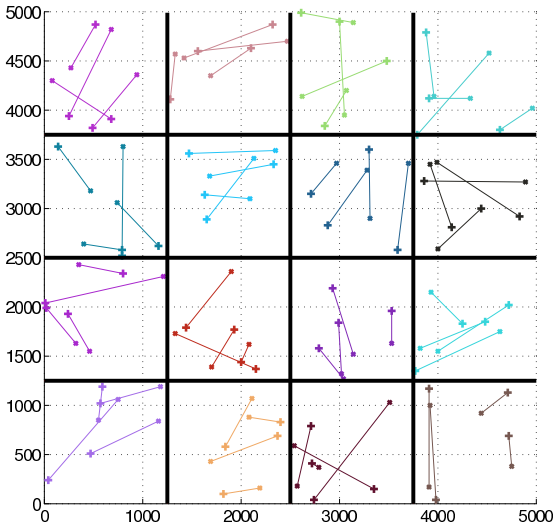


Fig. 3. A 5 km $\times$ 5 km square field divided into $K = 16$ cells. Nodes are positioned randomly inside each cells.

V. Simulation Results

This section provides numerical evaluations of all results presented in this paper. To implement these simulations, two scenarios are considered and reported in Figure 1 and Figure 3, respectively.

A. Numerical Validation

Theorems 4 and 5 allow the calculation of the fraction of time the system uses an NE and the average number of iterations needed before visiting the NE for the first time, as a function of several design parameters when the channel gains follow (12). The following shows that these results also hold under a more general formulation.

All experiments presented here are run on the scenario represented in Figure 1, with two different sets of parameters. The first set is composed of: $K = 3$, $C = 4$, $\epsilon = 0.02$ and $6 \leq Q \leq 10$; the second one is composed of $K = 4$, $C = 5$, $\epsilon = 0.02$ and $6 \leq Q \leq 10$. In the first experiment, the fraction of time the system uses an NE is estimated by running 10^7 iterations under two different channel models: the simple channels expressed in (12) and a channel power gain randomly drawn from a Rayleigh distribution. These results are summarized in Fig. 4. The dashed line and the continuous line correspond to the theoretical results with the first and the second set of parameters respectively. The results of the simulations are close to the lines for both channel models.

In the second experiment, the number of iterations needed to visit an NE for the first time is estimated and compared with the analytical results in Fig.5. Increasing the dimension of the action set, i.e., increasing C or Q, brings slower convergence rates since the algorithm requires more time to explore all the possibilities.

B. Convergence Nash Equilibrium

The following shows the effect of the enhancement on the stability and in the speed of the algorithm in reaching any

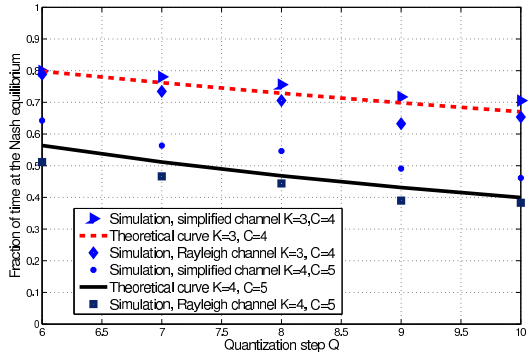


Fig. 4. Fraction of time the system is at the NE, with the TE learning algorithm, $\epsilon = 0.01$ and uniform probability distribution over the action set. Theoretical results are represented by the continuous lines, simulation results are represented by the markers for two sets of data and different channels: Rayleigh and the model in (12).

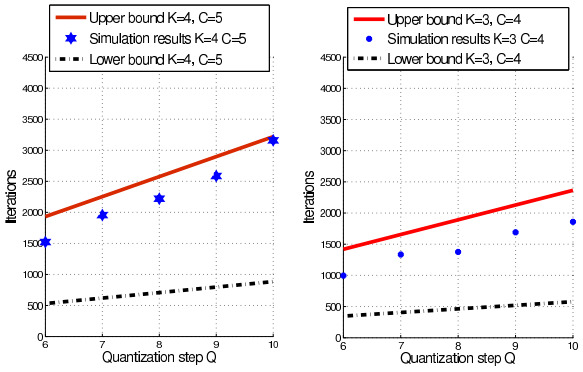


Fig. 5. Number of iterations needed for the algorithm to visit an NE for the first time. Simulations run with the standard TE, $\epsilon = 0.01$ and uniform probability distribution on the actions set. The continuous lines represent (17), the dashed lines represent (18).

stochastically stable point. A total of 10^4 iterations of TE are run with an underlying network as the one depicted in Figure 3, with $K = 4$ cells each populated with one link, $C = 4$ channels, $Q = 5$ power levels and a target SINR of $\Gamma = 10$ dB. In Fig. 6 the probability with which the TE algorithm selects an NE as a network action profile is plotted as a function of the experimentation probabilities ϵ_p and ϵ_c^{min}. Reducing the minimum experimentation probability on the channel sensibly decreases the instability of the system and thus increases the probability of the system of being at the NE. On the other hand, the stabilizing effect of reducing the experimentation probability on the power levels is balanced by the longer time that is needed for the system to reach an NE, as showed in Fig.7. In this figure, the number of iterations used by the TE learning algorithm to reach, for the first time, an NE is plotted as a function of the experimentation probabilities ϵ_p and $\epsilon_c^{\min}$. Note that, the number of iterations needed to reach for the first time an NE represents also a measure of the speed of the algorithm to reach again an NE, once it is left. From a real-system implementation point of view, it is also an estimation of the ability of the algorithm to react to network changes that modify the NE set, e.g. fading, shadowing, mobility, etc. By inspecting both plots, it appears that the experimentation

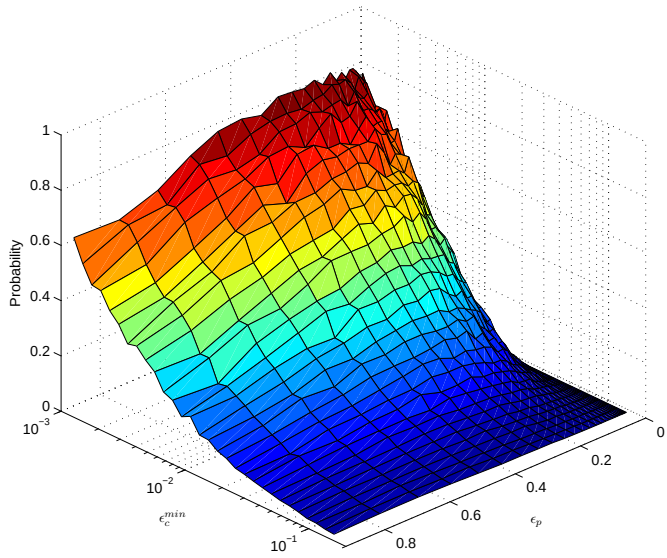


Fig. 6. The plot represents the probability of observing the TE learning algorithm selecting an action profile which is an NE as a function of ϵ_p and $\epsilon_c^{\min}$. The underlying network is composed of $K = 4$ cells, $L_k = 1$ links per cell, $C = 4$ channels and $Q = 5$ power levels. The $\epsilon_c^{\min}$ values are reported in logarithmic scale.

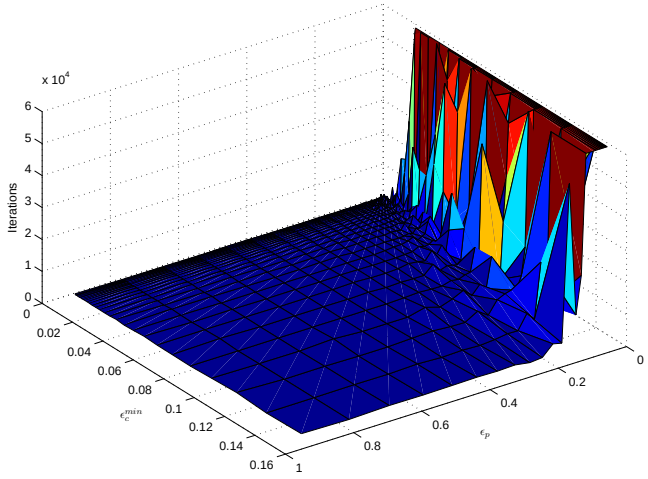


Fig. 7. The plot represents the number of iterations between $t = 0$ and the instant in which an NE is played for the first time as a function of ϵ_p and $\epsilon_c^{\min}$. The underlying network is the same as in Fig. 6.

frequency on the power levels should be relatively high, while the one on the channels should be relatively low with the exact optimal values depending on the other parameters of the network.

C. Performance Metrics

The following metrics are considered to evaluate the performance of the proposed algorithm:

- Average satisfaction (AS): The average number of times a link satisfies its SINR constraints.

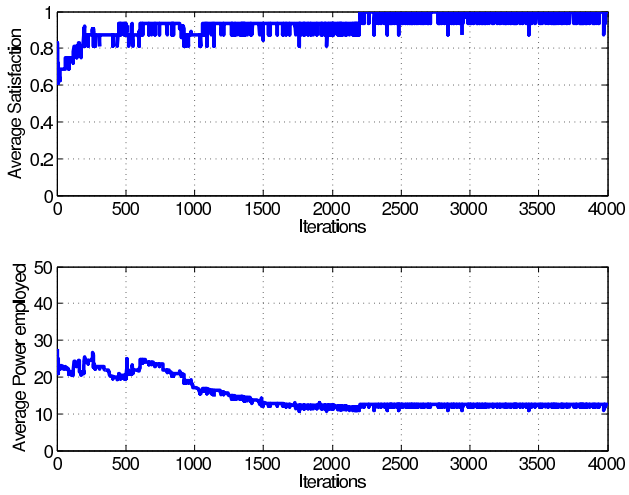


Fig. 8. The upper plot represents the AS, the lower plot represents the APC.

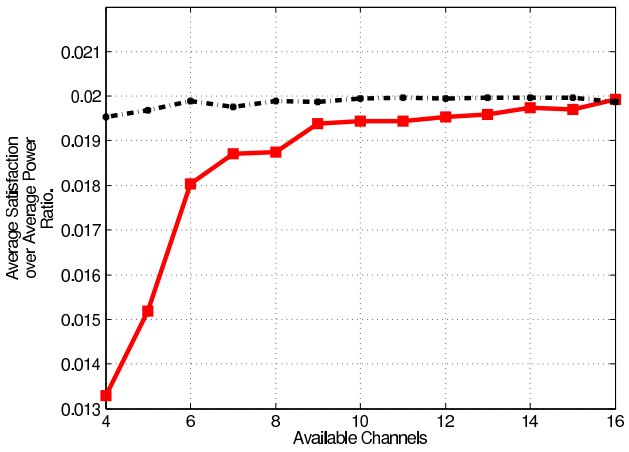


Fig. 9. Performance comparison between TE and the GBDCA in terms of average number of constraints satisfied over average used power. The dashed line is the performance of TE and the continuous line the one of GBDCA.

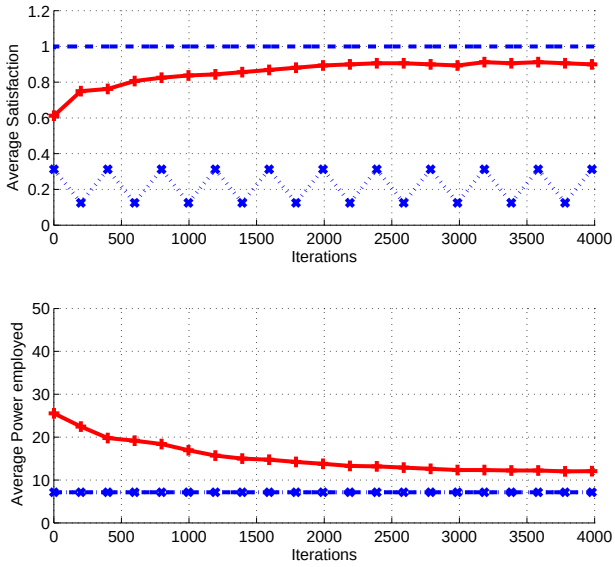


Fig. 10. Performance comparison between TE (red continuous line), the synchronous IWF (dotted line) and the global optimum (dashed line). We represent, in the upper plot, the AS, in the lower plot, the APC. We run 400 iterations of each algorithm on a network composed of $K = 16$ cells, each populated with one link, $C = 5$ channels, $Q = 5$ power levels, a maximum available power of $P_{\max} = 50$W.

- Average power consumption (APC): The average power used by the transmitters in a cell to achieve the corresponding satisfaction level.
- Average satisfaction over average used power ratio: This metric establishes the ability of the algorithm in satisfying the constraints with respect to the average power used.

The simulation scenario is represented in Fig. 3. Consider a static network composed of $K = 16$ cells each with $N_k = 4$ links, $C = 10$ channels, and the maximum power $P_{\max} = 50$W is quantized in $Q = 8$ logarithmic levels. The results are reported in Fig. 8, where the upper plot represents the AS while the lower plot shows the APC. The figure shows that the TE algorithm is able to drive the network to an almost full satisfaction by averagely employing only 10W. Note that, even though the first visit to an NE may happen quite late, the global performance at non-equilibrium states is high. This is due to the fact that the probability of playing an action grows with the social welfare of the action itself [32]. Second, Fig. 8 shows that even when an equilibrium is achieved, the system sometimes attempts to use sub-optimal action profiles. This is due to the stochastic nature of the TE learning algorithm. Note that there exist a natural tradeoff between the time needed to visit an NE and stability of such an equilibrium. In order to decrease the time needed to visit an NE, the experimentation probability needs to be large while, in order to improve the stability it needs to be small.

Furthermore, the TE learning algorithm is compared with the greedy based decentralized algorithm (GBDCA) described in [14]. Briefly, this algorithm solves the graph-coloring problem, by letting each CC detect the channel employed by its neighbors. If a CC detects that it is using a channel already occupied by one of its neighbors then it chooses randomly another channel among the free ones. If no channel is free, then it does not change its strategy. Since this algorithm does not consider a power allocation policy, its transmission power is set to $P_{\max}$. In this context, the GBDCA is compared with the TE learning algorithm when the quantization levels are reduced to $Q = 2$, i.e., an ON-OFF policy. The results, in terms of the ratio $\frac{\sum_{\ell \in \mathcal{L}} \mathbb{1}_{\{\xi_\ell > \Gamma\}}}{\sum_{k \in \mathcal{K}} p_k}$ are reported in Fig.9. The TE learning algorithm allows the cells that cannot satisfy their constraints to stop the transmission for a short period of time, which increases the efficiency.

The following compares the performance of the TE learning algorithm with the one of synchronous IWF and the global optimum. Consider $K = 16$ cells, $N_k = 1$ link per cell, $C = 5$ channels, $Q = 5$ power levels and a target SINR $\Gamma = 10$ dB. In the synchronous IWF each transmitter has full knowledge of the transmit channel state information; each transmitter may exploit multiple channels; the power allocation routine happen at the same instant for all transmitters; and each transmitter attempts to achieve a transmission rate equal to $\log_2(1 + \Gamma)$ with the minimum necessary power. The results

of the experiment are reported in Fig. 10.

The first figure reports the AS in the upper plot and the APC in the lower plot. In these plots, the dashed line represents the global optimum, the continuous red line the performance of TE algorithm and the dotted line the performance of the synchronous IWF. The action profiles chosen by the TE algorithm approach the global optimum both in terms of constraints satisfaction and in terms of power drain. The synchronous IWF, even though it is allowed to exploit a larger amount of information, is not able to select an action that satisfies the constraints for a large proportion of the links.

VI. Conclusion

In this work, strong connections between the solutions to a centralized network optimization problem and the Nash equilibria of a given game has been established via the design of the corresponding utility functions. More specifically, it has been shown that by properly choosing the utility function, it is possible to make a decentralized network to be stable at a global optimal operating point. More importantly, it has been shown that such equilibria can also be achieved by using learning algorithms following the paradigm of trial and error. The intuitions on the utility function design as well as the relevance of the trial and error learning are shown using the the scenario of a decentralized multi-channel ad hoc network. Using this network model, some heuristic enhancements have been also presented to improve the convergence of the algorithm and theoretical bounds on the time to reach an equilibrium are formally proved.

Appendix A
Proof of theorem 2

Proof: Consider two arbitrary NE $\boldsymbol{a}^*$ and $\boldsymbol{a}^+ \in \mathcal{A}_{\rm NE}$, such that $\sum_{\ell\in\mathcal{L}} \mathbb{1}_{\{\xi(\boldsymbol{a}^*)>\Gamma\}} = L^*$, $\sum_{\ell\in\mathcal{L}} \mathbb{1}_{\{\xi(\boldsymbol{a}^+)>\Gamma\}} = L^+$ with $L^* \geq L^+ + 1$. From Theorem 1, the stochastically stable points of the TE algorithm are the NE that maximize the social welfare W. Therefore, proving the thesis is equivalent to proving that $W(\boldsymbol{a}^*) > W(\boldsymbol{a}^+)$.

The social welfare associated with $\boldsymbol{a}^*$ using the utility in (3) is

$$\begin{aligned} W(\boldsymbol{a}^*) &= \sum_{k\in\mathcal{K}} u_k(\boldsymbol{a}^*) \\ &= \sum_{k\in\mathcal{K}} \frac{1}{1+\beta L_{\max}} \left(\varphi_k(\boldsymbol{a}^*) + \beta \sum_{\ell\in\mathcal{L}_k} \mathbb{1}_{\{\xi_i(\boldsymbol{a}^*)>\Gamma\}} \right) \\ &= \frac{1}{1+\beta L_{\max}} \left(\beta L^* + \sum_{k=1}^{K} \varphi_k(\boldsymbol{a}^*) \right). \end{aligned} \tag{20}$$

Since φ_k is a non-negative function, it holds that

$$W(\boldsymbol{a^*}) \geq \frac{\beta L^*}{1+\beta L_{\max}}. \tag{21}$$

Analogously, the social welfare associated with $\boldsymbol{a}^+$ is

$$\begin{aligned} W(\boldsymbol{a}^+) &= \sum_{k\in\mathcal{K}} u_k(\boldsymbol{a}^+) \\ &= \sum_{k\in\mathcal{K}} \frac{1}{1+\beta L_{\max}} \left(\varphi_k(\boldsymbol{a}^+) + \beta \sum_{\ell\in\mathcal{L}_k} \mathbb{1}_{\{\xi_i(\boldsymbol{a}^+)>\Gamma\}} \right) \\ &= \frac{1}{1+\beta L_{\max}} \left(\beta L^+ + \sum_{k=1}^{K} \varphi_k(\boldsymbol{a}^+) \right). \end{aligned} \tag{22}$$

By definition, $\forall\ \boldsymbol{a} \in \mathcal{A}^K$, and $\forall\ k \in \mathcal{K}$, $\varphi_k(\boldsymbol{a}) \leq 1$ and thus $W(\boldsymbol{a}^+) \leq \frac{\beta L^+ + K}{1+\beta L_{\max}}$. Then, using the assumption that $L^+ \leq L^* - 1$, it holds that

$$W(\boldsymbol{a}^+) \leq \frac{\beta L^* - \beta + K}{1+\beta L_{\max}}.$$

Therefore, from the assumption that $\beta > K$ it is possible to write

$$\frac{\beta L^* - \beta + K}{1+\beta L_{\max}} < \frac{\beta L^*}{1+\beta L_{\max}}, \tag{23}$$

thus, following the chain of inequalities, it holds that $W(\boldsymbol{a}^+) < W(\boldsymbol{a}^*)$. This concludes the proof. ■

Appendix B
Proof of theorem 3

Proof: From the assumptions of Theorem 3, the intersection between the set of NE $\mathcal{A}_{\rm NE}$ and the set of solutions of (1) $\mathcal{A}^\dagger$ is non empty, i.e., $\mathcal{A}_{\rm NE} \cap \mathcal{A}^\dagger \neq \emptyset$. Let $\boldsymbol{a^*} \in \mathcal{A}_{\rm NE} \cap \mathcal{A}^\dagger$ be an arbitrary element of the intersection and $L^* = \sum_{\ell\in\mathcal{L}} \mathbb{1}_{\{\xi(\boldsymbol{a}^*)>\Gamma\}}$ the number of links that satisfy their constraints. Since $\boldsymbol{a^*} \in \mathcal{A}^\dagger$ it results that $L^* = \max_{\boldsymbol{a}\in\mathcal{A}^K} \sum_{\ell\in\mathcal{L}} \mathbb{1}_{\{\xi(\boldsymbol{a})>\Gamma\}}$, i.e., L^* is the maximum number of links that can simultaneously satisfy their constraints. From Theorem 1, the set of the stochastically stable action profiles is $\mathcal{A}_{\rm TE} = \left\{ \boldsymbol{a}' \in \mathcal{A}^K : \boldsymbol{a}' \in \arg\max_{a\in\mathcal{A}_{\rm NE}} W(\boldsymbol{a}) \right\}$. Hence, proving the theorem is equivalent to prove that $\mathcal{A}_{\rm TE} \subseteq \left(\mathcal{A}_{\rm NE} \cap \mathcal{A}^\dagger\right)$. From its definition $\mathcal{A}_{\rm TE} \subseteq \mathcal{A}_{\rm NE}$, thus it remains to prove that $\mathcal{A}_{\rm TE} \subseteq \mathcal{A}^\dagger$.

Let $\mathcal{A}^\star \subseteq \mathcal{A}_{\rm NE}$ be the set of NE such that $\forall\ \boldsymbol{a} \in \mathcal{A}^\star$ $\sum_{\ell\in\mathcal{L}} \mathbb{1}_{\{\xi_\ell(\boldsymbol{a})>\Gamma\}} = L^*$. Then, it results that $\forall\ \boldsymbol{a}^+ \in \mathcal{A}^K \setminus \mathcal{A}^\star$ it hold that $\sum_{\ell\in\mathcal{L}} \mathbb{1}_{\left\{\xi_\ell(\boldsymbol{a}^+)>\Gamma\right\}} < L^*$. Thus, from Theorem 2 and the assumption that $\beta > K$, it holds that $W(\boldsymbol{a}^+) < W(\boldsymbol{a})$, $\forall\ \boldsymbol{a}^+ \in \left(\mathcal{A}^K \setminus \mathcal{A}^\star\right)$ and $\forall\ \boldsymbol{a} \in \mathcal{A}^\star$. Therefore the set of stochastically stable points can be expressed as $\mathcal{A}_{\rm TE} = \left\{ \boldsymbol{a}' \in \mathcal{A}^K : \boldsymbol{a}' \in \arg\max_{a\in\mathcal{A}^\star} W(\boldsymbol{a}) \right\}$. The social welfare of the action profiles on $\mathcal{A}^\star$ is:

$$W(\boldsymbol{a}) = \beta L^* + \sum_{k=1}^{K} \varphi_k(\boldsymbol{a}). \tag{24}$$

Therefore, $\arg\max_{\boldsymbol{a}\in\mathcal{A}^\star} W(\boldsymbol{a}) = \arg\max_{\boldsymbol{a}\in\mathcal{A}^\star} \sum_{k=1}^{K} \varphi_k(\boldsymbol{a})$. Thus, $\mathcal{A}_{\rm TE}$ is the set of the action profiles that satisfy the constraints for L^* links and maximizes the $\sum_{k=1}^{K} \varphi_k(\boldsymbol{a})$, hence $\mathcal{A}_{\rm TE} \subseteq \mathcal{A}^\dagger$.

This concludes our proof. ■

APPENDIX C
MARKOV CHAIN TRANSITION PROBABILITIES

A. Transition probability from an NE to a discontent state

The transition probability between an NE state and a state with one *discontent* player is denoted by $P(N,D)$. For the system to exit an NE, a player must pass from a *content* to a *discontent* state. This happens only in the following case: at time t player k experiments and during this experimentation k interferes with enough power to turn player l into watchful, at time $(t+1)$ player m experiments and during this experimentation m interferes turning l into *discontent*. The probability of at least one player experimenting in the system is given by: $P_\epsilon = 1-(1-\epsilon)^K$. By using the first two terms (reasonable since $\epsilon \ll 1$ implies $\epsilon^N \ll \epsilon^{(N-1)}$) of the binomial expansion $(1+(-\epsilon))^K = \sum_{k=0}^{K}\binom{K}{k}(-\epsilon)^k$ it holds that $P_\epsilon \simeq K\epsilon$. The probability that the player k disturbs another one, say l, is given by: (a) the probability of choosing an already occupied channel multiplied by (b) the probability of selecting a power level high enough. As a worst case scenario, assume that any power level greater than first quantization level is enough to create an intolerable level of interference. Thus, this probability is given by:

$$P_d = \underbrace{\frac{K-1}{C}}_{(a)}\underbrace{\frac{(Q-1)}{Q}}_{(b)}. \tag{25}$$

The probability that a player different from l experiments is $(K-1)\epsilon$, the probability of choosing the channel employed by l is $\frac{1}{C}$ and the probability of selecting a power level high enough is again given by (25). Therefore,

$$P(N,D) = K\epsilon\tfrac{(K-1)}{C}\tfrac{(Q-1)}{Q}(K-1)\epsilon\tfrac{1}{C}\tfrac{(Q-1)}{Q} \tag{26}$$

$$= \tfrac{K(K-1)^2\epsilon^2}{C^2}\left(\tfrac{Q-1}{Q}\right)^2. \tag{27}$$

B. Transition probability from discontent state to an NE

Here, we aim at evaluating $P(D,N)$, i.e., the transition probability between a state in which one player is *discontent* and a state in which all players are at the NE. Therefore, one player is performing a *noisy* search. Thus, the probability of immediately returning to an NE is given by: (a) the probability of selecting a free channel times (b) the probability of selecting enough power. Thus, we obtain:

$$P(D,N) = \underbrace{\frac{C-(K-1)}{C}}_{(a)}\underbrace{\frac{1}{Q}}_{(b)}. \tag{28}$$

C. Transition probability from a discontent state to a content state

The transition probability from a state with one *discontent* player to a state in which $K-k$ players are employing an individually optimal action is denoted by $P(D,C_{K-k})$. The *discontent* player selects a random action, then the probability of quitting the *discontent* state to a state in which only $(K-k)$ players are using one of their individually optimal actions depends on the acceptance function $F(u)$. Given (7) and for K large enough, the accepting probability can be approximated by $\epsilon^{F(u)} \approx 1$. When a player is *discontent*, it is possible for it to accept as a benchmark action the one that makes another player to change into a *discontent* mood. Then, the transition probability towards state C_{K-k} is given by the product of the probability of disturbing $(k-1)$ players that were at an NE before selecting a free channel or a channel used by a player that is not at an NE. The probability of colliding with $k-1$ players is given by

$$\frac{(K-1)}{C}\frac{(K-2)}{C}\frac{(K-3)}{C}\cdots\frac{(K-k+1)}{C} = \frac{(K-1)!}{C^{k-1}\,(K-k)!}, \tag{29}$$

while the probability of selecting a channel free or used by a player not using an individually optimal action is $\frac{C-(K-k)}{C}$. Therefore, the product is:

$$P(D,C_{K-k}) = \frac{1}{C^k}\frac{(K-1)!}{(K-k)!}\left(C-K+k\right). \tag{30}$$

D. Transition probability from C_{K-k} to C_{K-k+1}

The transition probability between a state in which $K-k$ players are using an individually optimal action and a state in which $K-k+1$ players are using an individually optimal action is denoted by $P\left(C_{K-k},C_{K-k+1}\right)$. Since no player is *discontent*, the transition happens through experimentation. To pass from a state in which $K-k$ players are using an individually optimal action to another one in which $K-k+1$ are doing the same, the following sequence of events must happen: at least one of the $K-k$ players experiments; it selects one of the available individually optimal actions; and it accepts the action. Thus, the transition probability is

$$P\left(C_{K-k},C_{K-k+1}\right) = \underbrace{(K-k)\epsilon}_{(a)}\underbrace{\frac{C-k}{CQ}}_{(b)}\underbrace{\epsilon^{G(\Delta u)}}_{(c)} \tag{31}$$

$$= (K-k)\frac{C-k}{CQ}\epsilon^{1+G(\Delta u)}. \tag{32}$$

APPENDIX D
PROOF OF THEOREM 4

Proof: With a standard Markov chain analysis, starting from state C_0, the expected number of iterations before reaching for the first time the NE is given by: $\bar{T}_{NE} = \sum_{k=0}^{K-1}\frac{1}{P(C_{K-k},C_{K-k+1})}$. Substituting, we obtain

$$\bar{T}_{NE} = \frac{CQ}{\epsilon^{(1+G(\Delta u))}}\sum_{k=0}^{K-1}\frac{1}{(K-k)\,(C-k)}$$

$$= \frac{CQ}{\epsilon^{(1+G(\Delta u))}\,(C-K)}\sum_{k=0}^{K-1}\left(\frac{1}{K-k}-\frac{1}{C-k}\right) \tag{33}$$

Given (6) and the fact that $\epsilon \ll 1$, the following approximation holds $\epsilon^{(1+G(\Delta u))} \approx \epsilon$. For the sake of simplicity, in the following, the pre-multiplying constant factor is omitted and define $m = K-k$. Thus, equation (33) can be written as

$$\sum_{m=1}^{K}\left(\frac{1}{m}-\frac{1}{C-K+m}\right). \tag{34}$$

It is known that $\sum_{m=1}^{K}\frac{1}{m} < 1 + \int_1^K \frac{1}{x}dx$ thus:

$$\sum_{m=1}^{K}\frac{1}{m} \leq \log(K) + 1. \tag{35}$$

It is also known that the harmonic sum is such that

$$\sum_{m=1}^{K}\frac{1}{m} \geq \log(K) + \gamma. \tag{36}$$

Consider that $\forall n \geq 1$, with $K \in \mathbb{N}$ and $A \in \mathbb{N}$,

$$\int_n^{K+1}\frac{1}{A+x}dx < \sum_{m=n}^{K}\frac{1}{A+m} < \int_{n-1}^{K}\frac{1}{A+x}dx, \tag{37}$$

and thus, for the second addend it holds that:

$$\sum_{m=1}^{K}\frac{1}{C-K+m} \leq \log\left(\frac{C}{C-K}\right), \tag{38}$$

$$\sum_{m=1}^{K}\frac{1}{C-K+m} \geq \log\left(\frac{C+1}{C-K+1}\right). \tag{39}$$

By joining together equation (35) with (38) and (36) with (39), and by reinserting the omitted multiplicative factor, we obtain the result, and this concludes the proof. ■

Appendix E
Proof of Theorem 5

Proof: The average fraction of time the system is at an NE can be expressed as $(1-\delta) = \frac{T_N}{T_{TOT}}$, where $\bar{T}_N$ is the expected time spent at an NE once it has been reached and by T_{TOT} the total time spent in all the states. Given the DTMC in Fig. 2, this can be expressed as

$$T_{TOT} = \bar{T}_N + T_{BNE}, \tag{40}$$

where $\bar{T}_{BNE}$ denotes the expected time between the instant the system leaves an NE and the instant it reaches it again. The expected number of time steps needed to leave the NE once reached is

$$\begin{aligned}\bar{T}_N = & \sum_{n=1}^{\infty} nP(NE,D)(1-P(NE,D))^{(n-1)} \\ = & -P(NE,D)\tfrac{d}{dP(NE,D)}\sum_{n=1}^{\infty}(1-P(NE,D))^n \\ = & -P(NE,D)\tfrac{d}{dP(NE,D)}\left(\tfrac{1}{P(NE,D)}\right) \\ = & \tfrac{1}{P(NE,D)}.\end{aligned}$$

Here, the well known equality $\sum_{n=1}^{\infty}x^n = \frac{x}{1-x}$ has been used and $\sum_{n=1}^{\infty}nx^{(n-1)} = \frac{d}{dx}\sum_{n=1}^{\infty}x^n$. Thus, it follows that

$$(1-\delta) = \frac{1}{1+P(NE,D)T_{BNE}}. \tag{41}$$

To evaluate T_{BNE}, the process is as follows. The starting state on the Markov chain is the state D. From here, it is possible to go back to the NE state without quitting the discontent state. To do this, the expected number of time steps needed is $T_{(D,NE)} = \sum_{n=1}^{\infty}nP(D,N)P(D,D)^{(n-1)}$. These equalities imply the following

$$T_{(D,NE)} = \frac{P(D,NE)}{(1-P(D,D))^2}, \tag{42}$$

where $P(D,D)$ is easily obtained by imposing the sum of the probabilities to be equal to 1:

$$P(D,D) = 1 - \left(P(D,NE) + \sum_{k=1}^{K}P(D,C_{K-k})\right). \tag{43}$$

On the other hand, it is possible to transit from the discontent state to a certain C_{K-k} state and the expected time steps needed to return to the NE starting from state C_{K-k} is denoted by $T_{CNE}(k)$. This quantity can be upper-bounded by using (37):

$$T_{CNE}(k) \leq \frac{CQ}{\epsilon^{1+G(\Delta u)}(C-K)}\left(\gamma + \log\left(\frac{K(C-k+1)}{C+1}\right)\right). \tag{44}$$

In the following, this upper bound is used as a close enough approximation of the true value. Moreover, given (6), and $\epsilon \ll 1$, it follows that $\epsilon^{1+G(\Delta u)} \approx \epsilon$. As consequence, the expected time T_{BNE} to return to an NE when the system deviates is given by:

$$T_{BNE} = T_{(D,NE)} + \sum_{k=1}^{K}P(D,C_{K-k})T_{CNE}(k). \tag{45}$$

This concludes the proof. ■

Acknowledgment

This research work was carried out in the framework of the CORASMA EDA Project B-0781-IAP4-GC.

References

[1] J. Hoydis, M. Kobayashi, and M. Debbah, "Green small-cell networks," *IEEE Vehicular Technology Magazine*, vol. 6, no. 1, pp. 37–43, Mar. 2011.

[2] W. Kiess and M. Mauve, "A survey on real-world implementations of mobile ad-hoc networks," *Ad Hoc Networks*, vol. 5, no. 3, pp. 324–339, Apr. 2007.

[3] L. Rose, S. Lasaulce, S. M. Perlaza, and M. Debbah, "Learning equilibria with partial information in decentralized wireless networks," *IEEE Communications Magazine*, vol. 49, no. 8, pp. 136–142, Aug. 2011.

[4] I. Macaluso, L. D. Silva, and L. Doyle, "Learning Nash equilibria in distributed channel selection for frequency-agile radios," in *Proc. Workshop on Artificial Intelligence for Telecommunications and Sensors Networks (WAITS)*, Montpellier, France, Apr. 2012.

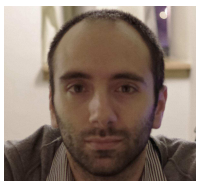
[5] Y. Xu, A. Anpalagan, Q. Wu, L. Shen, Z. Gao, and J. Wang, "Decision-theoretic distributed channel selection for opportunistic spectrum access: Strategies, challenges and solutions," *IEEE Communications Surveys Tutorials*, vol. 1, no. 99, pp. 1–25, Apr. 2013.

[6] D. Gesbert, S. Kiani, A. Gjendemsjo, and G. ien, "Adaptation, coordination, and distributed resource allocation in interference-limited wireless networks," *Proc. the IEEE*, vol. 95, no. 12, pp. 2393–2409, Dec. 2007.

[7] Y. Xu, J. Wang, Q. Wu, A. Anpalagan, and Y.-D. Yao, "Opportunistic spectrum access in cognitive radio networks: Global optimization using local interaction games," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 2, pp. 180–194, Apr. 2012.

[8] C. Tekin, M. Liu, R. Southwell, J. Huang, and S. Ahmad, "Atomic congestion games on graphs and their applications in networking," *IEEE ACM Transactions on Networking*, vol. 20, no. 5, pp. 1541–1552, Oct. 2012.

[9] R. D. Yates, D. Tse, and Z. Li, "Secret communication on interference channels," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Toronto, Canada, Jul. 2008.
[10] R. Berry and D. Tse, "Shannon meets nash on the interference channel," *IEEE Transactions on Information Theory*, vol. 57, no. 5, pp. 2821–2836, May 2011.
[11] S. M. Perlaza, R. Tandon, H. V. Poor, and Z. Han, "The Nash equilibrium region of the linear deterministic interference channel with feedback," in *Proc. 50th Annual Allerton Conference on Communications, Control, and Computing*, Monticello, IL, Oct. 2012.
[12] H. P. Young, "Learning by trial and error," *Games and Economic Behavior*, vol. 65, no. 2, pp. 626–643, Mar. 2009.
[13] D. Fudenberg and J. Tirole, *Game Theory*. Cambridge, MA, USA: The MIT Press, 1991.
[14] T.-C. Hou and T.-J. Tsai, "On the cluster based dynamic channel assignment for multihop ad hoc networks," *Journal of Communications and Networks*, vol. 4, no. 1, pp. 40–47, Mar. 2002.
[15] J.-S. Pang, G. Scutari, D. P. Palomar, and F. Facchinei, "Design of cognitive radio systems under temperature-interference constraints: A variational inequality approach," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3251–3271, Jun. 2010.
[16] V. L. Nir and B. Scheers, "Improved coexistence between multiple cognitive tactical radio networks by an expert rule based on sub-channel selection," in *Proc. Wireless Innovation Forum European Conference on Communications Technologies and Software Defined Radio SDR'11 (WInnComm-Europe)*, Brussels, Belgium, Jun. 2011.
[17] V. Le Nir and B. Scheers, "Autonomous dynamic spectrum management for coexistence of multiple cognitive tactical radio networks," in *Proc. IEEE Fifth International Conference on Cognitive Radio Oriented Wireless Networks Communications (CROWNCOM)*, Cannes, France, Jun. 2010.
[18] Y. Xu, J. Wang, Q. Wu, A. Anpalagan, and Y.-D. Yao, "Opportunistic spectrum access in unknown dynamic environment: A game-theoretic stochastic learning solution," *IEEE Transactions on Wireless Communications,*, vol. 11, no. 4, pp. 1380–1391, Apr. 2012.
[19] D. Niyato and E. Hossain, "Cognitive radio for next-generation wireless networks: An approach to opportunistic channel selection in IEEE 802.11-based wireless mesh," *IEEE Transactions on Wireless Communications*, vol. 16, no. 1, pp. 46–54, Feb. 2009.
[20] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5667–5681, Nov. 2010.
[21] P. Sastry, V. Phansalkar, and M. Thathachar, "Decentralized learning of Nash equilibria in multi-person stochastic games with incomplete information," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 24, no. 5, pp. 769–777, May 1994.
[22] V. Belmega, S. Lasaulce, M. Debbah, and A. Hjorungnes, "Learning Distributed Power Allocation Policies in MIMO Channels," in *Proc. European Signal Processing Conference (EUSIPCO)*, Aalborg, Denmark, Aug. 2010.
[23] Y. Xing and R. Chandramouli, "Stochastic learning solution for distributed discrete power control game in wireless data networks," *IEEE/ACM Transactions on Networking*, vol. 16, no. 4, pp. 932–944, Aug. 2008.
[24] O. Popescu and C. Rose, "Water filling may not good neighbors make," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, San Francisco, CA, USA, Dec. 2003.
[25] L. Rose, S. M. Perlaza, and M. Debbah, "On the Nash equilibria in decentralized parallel interference channels," in *Proc. IEEE Workshop on Game Theory and Resource Allocation for 4G*, Kyoto, Japan, Jun. 2011.
[26] E. Altman, V. Kumar, and H. Kameda, "A Braess type paradox in power control over interference channels," in *Proc. 6th Intl. Symp. on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT)*, Berlin, Germany, Apr. 2008.
[27] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Transactions on Communications*, vol. 54, no. 7, pp. 1310–1322, Jul. 2006.
[28] R. Cendrillon, J. Huang, M. Chiang, and M. Moonen, "Autonomous spectrum balancing for digital subscriber lines," *IEEE Transactions on Signal Processing*, vol. 55, no. 8, pp. 4241–4257, Aug. 2007.
[29] R. Cendrillon, W. Yu, M. Moonen, J. Verlinden, and T. Bostoen, "Optimal multiuser spectrum balancing for digital subscriber lines," *IEEE Transactions on Communications*, vol. 54, no. 5, pp. 922–933, May 2006.
[30] M. Bennis, S. M. Perlaza, P. Blasco, Z. Han, and H. V. Poor, "Self-organization in small cell networks: A reinforcement learning approach," *IEEE Transactions on Wireless Communications*, vol. 12, no. 7, pp. 3202–3212, Jun. 2013.
[31] H. P. Young, *Strategic Learning and Its Limits (Arne Ryde Memorial Lectures Sereis)*. Oxford University Press, USA, 2004.
[32] B. S. Pradelski and H. P. Young, "Learning efficient Nash equilibria in distributed systems," Tech. Rep., Sep. 2010.
[33] L. Rose, S. M. Perlaza, M. Debbah, and C. L. Martret, "Distributed power allocation with SINR constraints using trial and error learning," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, Paris, France, Apr. 2012.
[34] J. F. Nash, "Equilibrium points in n-person games," *Proc. the National Academy of Sciences of the United States of America*, vol. 36, no. 1, pp. 48–49, 1950.



Luca Rose (S'10) was born in Pisa, Italy. He received his B.Sc. degree in telecommunications engineering from the University of Pisa, Italy. In 2009 he received his M.Sc. degree with honours from the same university. Since 2010, he has been attending his PhD studies on Game theory and Learning theory as a member of the Alcatel-Lucent chair in flexible radio in Supélec, sponsored by Thales Communications & Security. Furthermore, he holds a position as research engineer in Thales in the context of the project CORASMA (COgnitive RAdio for dynamic Spectrum MAnagement). His interests span from the field of software radio to the field of Game theory and Learning theory, small cells, and distributed resource allocation.

Samir M. Perlaza (M'13) received the B.Sc. degree from Universidad del Cauca, Popayán, Colombia, in 2005 and the M.Sc. and Ph.D. degrees from École Nationale Supérieure des Télécommunications (Telecom ParisTech), Paris, France, in 2008 and 2011, respectively. From 2008 to 2011, he held a position as a Research Engineer at France Télécom (Orange Labs, Paris, France) and during the second half of 2011 he was with the Alcatel Lucent Chair in Flexible Radio, Gif-sur-Yvette, France. Since 2012, he is a Post-Doctoral Research Associate in the Department of Electrical Engineering at Princeton University, Princeton, N.J. His research interests lie in the overlap of signal processing, information theory, game theory and wireless communications. Dr. Perlaza was a recipient of an Alβan Fellowship of the European Union in 2006 and the Best Student Paper Award in Crowncom in 2009.

Christophe J. Le Martret (SM'06) was born in Rennes, France, on March 12, 1963. He received the Ph.D. degree in Signal Processing and Communications from l'Universit de Rennes 1, Rennes, France, in 1990 and the HDR (Habilitation Diriger des Recherches) degree in 2010 and qualification for full professorship in 2012. From 1991 to 1995 he was with the CESTA, Bruz, France, and from 1996 to 2002 with the Centre d'lectronique de L'ARmement (CELAR). In 2002 he joined the Signal Processing and Multimedia Department at Thales Communications France. He was a Visiting Researcher at the SpinCom Laboratory, Department of ECE, University of Minnesota, MN, from 1999 to 2000.

He holds a Thales Expert position since 2007 in the field of wireless communications. Its current research activities cover radio access design with cross-layer optimization for ad hoc mobile networks, and cognitive radio.

Mérouane Debbah (SM'08) entered the Ecole Normale Supérieure de Cachan (France) in 1996 where he received his M.Sc and Ph.D. degrees respectively. He worked for Motorola Labs (Saclay, France) from 1999-2002 and the Vienna Research Center for Telecommunications (Vienna, Austria) until 2003. He then joined the Mobile Communications department of the Institut Eurecom (Sophia Antipolis, France) as an Assistant Professor until 2007. He is now a Full Professor at Supelec (Gif-sur-Yvette, France), holder of the Alcatel-Lucent Chair on Flexible Radio and a recipient of the ERC starting grant MORE (Advanced Mathematical Tools for Complex Network Engineering). His research interests are in information theory, signal processing and wireless communications. He is a senior area editor for IEEE Transactions on Signal Processing and an Associate Editor in Chief of the journal Random Matrix: Theory and Applications. Mérouane Debbah is the recipient of the "Mario Boella" award in 2005, the 2007 General Symposium IEEE GLOBECOM best paper award, the Wi-Opt 2009 best paper award, the 2010 Newcom++ best paper award, the WUN CogCom Best Paper 2012 and 2013 Award as well as the Valuetools 2007, Valuetools 2008, Valuetools 2012 and CrownCom2009 best student paper awards. He is a WWRF fellow and an elected member of the academic senate of Paris-Saclay. In 2011, he received the IEEE Glavieux Prize Award. He is the co-founder of Ximinds.